

## DOCUMENT RESUME

ED 228 307

TM 830 229

AUTHOR Smith, Richard M.  
TITLE Test Fairness Is a Personal Issue!  
PUB DATE Apr 83  
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).  
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)  
EDRS PRICE MF01/PC02 Plus Postage.  
DESCRIPTORS Academic Ability; Error of Measurement; Error Patterns; Higher Education; \*Individual Differences; Latent Trait Theory; \*Measurement Techniques; Psychometrics; Racial Bias; \*Scores; Sex Bias; \*Statistical Bias; \*Test Bias; Testing Problems; \*Test Items  
IDENTIFIERS New Jersey College Basic Skills Placement Test; \*Rasch Model

## ABSTRACT

Previous studies of test item bias have investigated how different groups of examinees perform differently on a given set of items. These studies imply that examinees should be treated in a certain way because they are of a particular sex or race rather than as individuals in their own right, but it is unrealistic and unfair to assume such an action is appropriate for every member of a group. Recent developments in psychometrics make it possible to consider each person, regardless of race or sex, as a unique population measureable only on those items which are "fair" for that individual. This study presents techniques for assessing whether and to what degree an individual was measured unfairly by a set of items. These techniques were applied to a sample from the statewide administration of the New Jersey College Basic Skills Test. The analysis determined the number of examinees in various subgroups who were not measured fairly and identified items which caused particular trouble.  
(Author)

\*\*\*\*\*  
\* Reproductions supplied by EDRS are the best that can be made \*  
\* from the original document. \*  
\*\*\*\*\*

U.S. DEPARTMENT OF EDUCATION  
NATIONAL INSTITUTE OF EDUCATION  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TEST FAIRNESS IS A PERSONAL ISSUE!

"PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

R. M. Smith

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)."

Richard M. Smith

American College Testing Program

April, 1983

Paper presented at the annual meeting of the National  
Council on Measurement in Education  
Montreal, April, 1983

ED228307

TM 830229

#### ABSTRACT

---

Previous studies of test item bias have investigated how different groups of examinees perform differently on a given set of items. These studies imply that examinees should be treated in a certain way because they are of a particular sex or race rather than as individuals in their own right, but it is unrealistic and unfair to assume such an action is appropriate for every member of a group. Recent developments in psychometrics make it possible to consider each person, regardless of race or sex, as a unique population measureable only on those items which are "fair" for that individual.

This study presents techniques for assessing whether and to what degree an individual was measured unfairly by a set of items. These techniques were applied to a sample from the statewide administration of the New Jersey College Basic Skills Test. The analysis determined the number of examinees in various subgroups who were not measured fairly and identified items which caused particular trouble.

### TEST FAIRNESS IS A PERSONAL ISSUE! \*

Scores which depend on knowing such things as a person's race or sex for their "correct" interpretation cannot provide a "fair" basis for the decisions required in education. The objective comparisons necessary for monitoring progress, for making decisions about competence, and for evaluating program effectiveness depend on unbiased measurements of each person's standing. For this it is necessary that differences in measures be due to differences in individual ability and not to bias in the measuring instrument (Wright, 1977, 1980).

The first attempts at dealing with bias in testing were directed at questions like "Is this test biased against a particular population of potential examinees?" This approach assumed that any bias was a property of tests and that the same adjustment was appropriate for each member of the specified population. The approach has been refined to deal with bias in individual items, but the treatment of examinees has remained at the group level.

In order for a measurement procedure to be fair it must be possible to keep the measure and its interpretation separate from considerations of all other characteristics of the person, including the groups to which he might belong. This condition

\* The author is indebted to Benjamin D. Wright for his many helpful comments and suggestions.

is not only necessary for the objective measurement of persons, it is also the necessary precondition for any study of the characteristics of persons or items which interfere with measurement as will be discussed more fully below.

In view of what is now possible with latent trait methods, the most significant shortcoming of previous work in test bias is its reliance on population-based corrections. If the aim is to treat each individual fairly, then it is unreasonable to assume that all the complexities and variations in cultures can be isolated in terms of a handful of demographic classifications. Because a set of items performs acceptably for almost everyone in a group does not mean that every member of the group can be adequately described by their performance on every one of those items. In addition, it is possible that some members of a group can be accurately characterized by their scores on a set of items even though some members cannot be.

Determining the extent to which an item is biased against a particular group does not mean that there is a single, simple adjustment that applies equally to each of the members of that group. It is still necessary to examine each person's individual responses to a set of items for any bias which is present for him, regardless of his group membership. A possible explanation for why so few studies have discovered significant biases against any group may be that when there is bias only some of the individuals in the group are strongly affected by

it.

The major thrust of earlier research has been to develop methods for constructing "culture-fair tests." It was hoped that this approach would produce tests which are appropriate for measuring anyone, but this is surely an unreachable goal. Using techniques derived from those described by Draba (1977) it should be possible, however, to develop "item banks" free of gross injustices and capable of defining a general variable that is internally consistent for most groups of interest. Even this, however, can never guarantee that every item will always be equally appropriate for every person regardless of what language they speak, how well they hear, how tired or sick they may be, or which hand they write with. If the goal is to obtain a "culture-free measure" for each person, it may be necessary to drop some items for some persons. This is because, while these items may work perfectly well for most persons, they have not worked this time for this person. The "test-freed" measurement of the Rasch model makes this individualized item editing possible and practical. The resulting "tailored" measure can then be interpreted and used regardless of the person's other attributes.

#### Rasch Person Analysis

The fundamental element for personal bias analysis is the interaction between the person and a single item. The Rasch model for dichotomously scored items gives the expected outcome

of this encounter. The probability that a person with ability "b" will succeed on an item with difficulty "d" is

$$p = \text{Prob}(X=1) = \exp(b-d)/(1 + \exp(b-d)). \quad (1)$$

This probability will be 0.5 when  $b = d$ ; it will approach 1 as  $b$  becomes greater than  $d$ ; and it will approach 0 as  $b$  becomes less than  $d$ .

The observed outcome of the encounter is either that the person answered the item correctly ( $x = 1$ ) or incorrectly ( $x = 0$ ). An index of the reasonableness of this response given the model is provided by the difference between the observed and predicted responses:

$$y = x - p. \quad (2)$$

Since  $y = (1-p)$  if the response is correct and  $y = -p$  if the response is incorrect, this can be interpreted as the significance level of the null hypothesis that this observation fits the model.

The validity of this approach is based on two propositions:

- (a) The items have been adequately calibrated for a relevant population.
- (b) The majority of the items have worked adequately for this person.

If the first is not true, then the items do not define a variable. If the second is not true, then there is insufficient information to accurately estimate this person's ability.

Both of these propositions are testable from the data. The

adequacy of the item calibration can be evaluated by the item fit statistics. The adequacy of the person's measure can be evaluated by the person fit statistics. A complete description of the Rasch person fit statistics and their type I error rates are found in Smith (1982).

There is too little information in a single observation, however, to arrive at any general conclusions. A more powerful analysis depends on prior information about persons and items that can be used to form clusters of items. These clusters can be based on any attribute of the items that might interact with examinee characteristics in any way which might confound the measurement of the examinee's ability.

For example, it is often thought that arithmetic knowledge can be measured equally well with items using numbers only as with items expressed in words. On the other hand, It is hypothesized that the verbally-loaded arithmetic items are unfair to persons whose expression of arithmetic knowledge is inhibited by interference from the extraneous reading variable, thus making the item appear to be more difficult for that person than it really is. Arithmetic items without words and items with words might, therefore, be chosen as two item sets in order to test the effect of verbal skills on the difficulty of arithmetic items.

The "test-freeness" of the Rasch model can be used to assess the extent of the "between item set" agreement for each

person. Thus, the person's ability can be estimated from each set without being influenced by the difficulties of the particular items involved. If all the estimates are statistically equivalent, it is safe to conclude that the person was fairly measured.

It is against this background that the analysis of real test data will be performed. This analysis is based on data collected for the New Jersey College Basic Skills Test during 1979. For this analysis the corrected unweighted fit statistics (Smith, 1982) will be used in order to assess the frequency of measurement disturbances in real test data.

Previous research (Mead, 1976; Wright & Mead, 1977; Wright, 1977; Wright & Stone, 1979; Smith & Mitchell, 1979, 1980; Mead & Kreines, 1980) suggests that a variety of test-taking skills and attitudes which influence achievement test scores can be identified in a systematic fashion through the analysis of individual fit statistics. The most common test-taking deficiencies are slow start-up, excessive carefulness, carelessness, guessing to complete, and random guessing.

#### Description of the Study

The New Jersey College Basic Skills Test was administered to 47,725 members of the Fall 1979 class of students entering public institutions of higher education. The majority of that class was tested between February and June of their senior year

in high school. The test is composed of six sections, five composed of multiple choice items and one consisting of a 20 minute essay. The two multiple choice sections used in this study were the Logical Relationships and Computation sub-tests. Each of the items on the tests is classified according to item content. The Logical Relationships test has four clusters: Categorizing Ideas, Connectives, Analogies, and Principles of Organization. The Computation test has four clusters: Operations with Whole Numbers, Operations with Fractions, Operations with Decimals, and Simple Applications.

In addition to the six test sections each student is requested to complete a personal information section. In 1979 this information included sex, date of birth, high school, full- or part-time status at college, year of graduation, whether English was the individual's best language, number of years of English and mathematics studied in high school, and a self-evaluation of written expression and mathematical ability.

Utilizing the data that were available on one administration of the New Jersey College Basic Skills Placement Test the following two-fold analysis was undertaken.

1) Identification of Demographically Biased Items

Items which exhibit demographic bias were identified using the technique proposed by Draba (1977) which compares item difficulties from calibrations on

different sub-populations. The identification of the sub-populations to be used for the detection of this kind of bias were based on the background information collected with each test. Random samples of 2000 individuals were drawn, when possible, for each subpopulation in the study and a bias analysis done for the Logical Relationships and Computation tests. With this analysis it was possible to identify items that exhibit sex bias as well as native language bias.

## 2) Personal Bias Analysis

On the basis of the various sets of biased items which were identified and the test-taking and test specific factors which can confound measurement, a complete person analysis was performed for random samples of 250 individuals from each cell of the design, i.e., males and females for whom English is their best language and for whom English is not their best language.

This analysis was completed for each of the two objective tests mentioned above. It identified the extent to which confounding factors were present in the data, and the magnitude of the impact of these factors on individual ability estimates. The analysis also suggested the nature and scope of remedial action necessary to remove the effects of the confounding

factors from each ability estimate.

Several item calibration samples were drawn from the 47,725 individuals who took the New Jersey College Basic Skills Test between May, 1979, and February, 1980. The first was a spaced random sample of 3151 examinees. Their responses determined the overall item difficulties for the two tests that were used in the person analysis program. This sample was drawn without regard to biographical information and is representative of the entire population. The BICAL program (Wright, Mead and Bell, 1979) was used throughout this study to estimate the item difficulties.

Subsequently four stratified random samples were drawn for the identification of sets of items that were differentially familiar for the subpopulations identified earlier. A sample of 2000 males who indicated English was their best language was drawn, as well as a sample of 2000 females who indicated English was their best language. Since there were only 842 males and 1017 females in the total population who indicated that English was not their best language the entirety of both subpopulations were used to determine the sets of differentially familiar items.

In order to determine sets of items that favored individuals on the basis of their best language, it was necessary to compare the item difficulties estimated from different subpopulations. A total of 1000 males and 1000

females who reported that English was their best language were combined for this calibration. All 842 male and 1017 female examinees who reported that English was not their best language were included in the second calibration sample. The item difficulties from these two calibration samples for the two tests were then plotted to determine which items differentially favored one or the other subpopulation.

In order to determine items that favored individuals on the basis of their sex, BICAL was used to estimate the item difficulties for the four calibration samples described above: 2000 males who regarded English as their best language; 2000 females who regarded English as their best language; 842 males who reported English was not their best language; and 1017 females who reported that English was not their best language. The item difficulties for the male/English best language sample and the female/English best language sample were plotted against each other. The item difficulties for the male/English not best language sample and the female/English not best language sample were also plotted.

A two standard error band was constructed about the identity line on each scaled plot. These bands were determined by the standard error of estimate for items located at various points along the identity line. Items that lie above or below these bands were considered to be differentially familiar to a particular subpopulation. See Wright and Stone (1979) for a

complete discussion of the procedure for determining the error band.

The analysis of the Logical Relationships test for sex bias is a replication of a study first reported in Smith (1981).

### RESULTS

#### Identification of Item Sets

A complete breakdown of the item type by language group favored for the Logical Relationships test is presented in Table 1.

TABLE 1  
DIFFERENTIALLY FAMILIAR ITEMS BY ITEM TYPE  
LOGICAL RELATIONSHIPS TEST

| Item Cluster<br>(item no.)          | Favor<br>English<br>Best Language | Neutral | Favor<br>English not<br>Best Language |
|-------------------------------------|-----------------------------------|---------|---------------------------------------|
| I (1-12)<br>Categorizing ideas      | 4                                 | 7       | 1                                     |
| II (13-25)<br>Connectives           | 0                                 | 7       | 6                                     |
| III (26-37)<br>Analogies            | 9                                 | 1       | 2                                     |
| IV (38-50)<br>Prin. of organization | 4                                 | 4       | 4                                     |

It is clear from this analysis that there is an item type by language interaction. One of the item types strongly favors and another moderately favors examinees who report that English is

their best language. One of the item types strongly favors examinees who report that English is not their best language. One of the item types favors neither group.

In order to demonstrate the relationship between items identified as differentially familiar to either sex and the stereotypical view of male or female-oriented activities, a list of the words used in such items on the Logical Relationship test is presented in Table 2.

TABLE 2  
WORDS USED IN DIFFERENTIALLY FAMILIAR ITEMS  
LOGICAL RELATIONSHIPS TEST

| Language<br>favored | Male                                                                                                                                                                              | Female                                                                                                                                                          |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Both                | Tanker<br>Yacht<br>Merchant ship<br>Telephone company<br>Electric company<br>Public utility<br>Weapon<br>Bayonet<br>Agreement<br>Treaty<br>Command<br>Obedience<br>Dent<br>Fender | Bustle<br>Dawdle<br>Disease<br>Health<br>Scandal<br>Reputation<br>Sherbet<br>Cream puff<br>Dessert<br>Contentment<br>Well-being<br>Flower<br>Corsage<br>Blossom |
| English<br>Best     | Lumber<br>Plank<br>American rock music                                                                                                                                            | Dental floss<br>Astrology                                                                                                                                       |

A complete breakdown of the item content by sex favored

for the items on the Computation test is presented in Table 3. In the subsequent person analysis the whole number classification shown in the table was dropped due to the low number of items. The native language group analysis indicated that only four Computation items favored native English speakers and only five favored non-English speakers. Since both of these classifications had relatively few items, the entire language group analysis was dropped for the Computation test.

TABLE 3  
DIFFERENTIALLY FAMILIAR ITEMS BY ITEM TYPE  
COMPUTATION TEST

| Sex     | Whole<br>Numbers | Fractions | Decimals | Application |
|---------|------------------|-----------|----------|-------------|
| Males   | 0                | 1         | 3        | 2           |
| Neutral | 1                | 6         | 5        | 5           |
| Females | 1                | 3         | 3        | 0           |

#### Person Analysis

To ascertain the degree to which measurement disturbances interfere with the measurement of individuals on the Logical Relationships and the Computation test, a sample of 250 examinees was drawn from each of the eight calibration samples. As described earlier, four single sex samples were drawn - 250 males and 250 females who reported that English was their best

language, as well as 250 males and 250 females who reported that English was not their best language were drawn for each test.

These samples were then analyzed with the PARM program (Smith, 1982b). The presence of measurement disturbances was detected by the use of three corrected unweighted fit statistics - total, between, and within. The determination of the most plausible measurement disturbance is based on the examination of the response pattern in a number of different configurations (Smith 1981b).

The determination of the item sets for the between fit statistic was based on the following: the characteristics of the items (item type, content, and position); and the calibrations, reported above, that identifying items differentially familiar to males and females and to examinees reporting that English was or was not their best language.

The basic types of measurement disturbances detected in this analysis are presented below.

#### Items Grouped by Content

- 1) Content/person interaction - The examinee scored unusually high or low on one or two item sets when compared with the remaining item sets.

#### Items Grouped by Difficulty

- 2) Guessing - A low ability examinee answered an approximately equal number of items of each difficulty correctly, or a low ability individual answered a surprising number of difficult items correctly.

- 3) Carelessness - A high ability individual missed a number of very easy items.

Items Grouped by Position on Test

- 4) Plodding - An examinee worked slowly but accurately. This usually resulted in a string of mostly correct responses followed by a string of omits.
- 5) Startup - An examinee did poorly on the items at the beginning of the test, possibly due to nervousness. Once this was overcome the examinee then worked at a uniformly higher ability.
- 6) Gave up - An examinee began the test but stopped working before the test was 1/2 over.

Items Grouped by Differential Familiarity

- 7) Sex - An examinee scored unusually high or low on an item set that should have been easier or harder based on the performance of other examinees of that sex.
- 8) Language - An examinee scored unusually high or low on an item set that should have been easier or harder based on the performance of other examinees who favor that language.

The number of examinees whose response patterns to the Logical Relationships test fit the previously described patterns is presented in Table 4. The same information is presented for the Computation Test in Table 5.

TABLE 4  
FREQUENCY OF MEASUREMENT DISTURBANCES BY SEX  
AND NATIVE LANGUAGE FOR THE LOGICAL RELATIONSHIPS TEST

| Type of<br>Measurement<br>Disturbance | Males<br>English<br>Best | Females<br>English<br>Best | Males<br>English<br>Not Best | Females<br>English<br>Not Best |
|---------------------------------------|--------------------------|----------------------------|------------------------------|--------------------------------|
| Carelessness                          | 8                        | 13                         | 8                            | 15                             |
| Content                               | 17                       | 18                         | 13                           | 15                             |
| Gave up                               | 2                        | 2                          | 0                            | 0                              |
| Guessing                              | 1                        | 4                          | 12                           | 2                              |
| Language                              | 2                        | 5                          | 12                           | 8                              |
| Plodding                              | 1                        | 4                          | 1                            | 5                              |
| Sex                                   | 10                       | 8                          | 6                            | 7                              |
| Startup                               | 0                        | 1                          | 2                            | 2                              |
| Unknown                               | 4                        | 4                          | 5                            | 9                              |
| Total                                 | 45                       | 59                         | 59                           | 63                             |
| Percent<br>of Sample                  | 18.0                     | 23.6                       | 23.6                         | 25.2                           |

Several patterns are apparent here. Overall, there is a higher frequency of measurement disturbances on the Computation test than on the Logical Relationships test. In particular, there appears to be much more guessing on the Computation test. This is understandable since the Logical Relationships test is the easiest in the battery with an average percent correct of 78%. The Computation test has an average percent correct of only 67%. There was also very little evidence of examinees giving up. This is attributable to the fact that both of these tests were relatively easy. It is usually assumed that the more difficult the test the more likely an examinee is to guess.

TABLE 5  
FREQUENCY OF MEASUREMENT DISTURBANCE BY SEX  
AND NATIVE LANGUAGE FOR THE COMPUTATION TEST

| Type of Measurement Disturbance | Males English Best | Females English Best | Males English Not Best | Females English Not Best |
|---------------------------------|--------------------|----------------------|------------------------|--------------------------|
| Carelessness                    | 15                 | 6                    | 16                     | 8                        |
| Content                         | 23                 | 33                   | 28                     | 25                       |
| Gave up                         | 1                  | 3                    | 4                      | 3                        |
| Guessing                        | 12                 | 16                   | 18                     | 18                       |
| Plodding                        | 0                  | 1                    | 0                      | 1                        |
| Sex                             | 11                 | 4                    | 11                     | 14                       |
| Startup                         | 2                  | 1                    | 4                      | 0                        |
| Unknown                         | 5                  | 6                    | 5                      | 1                        |
| Total                           | 69                 | 70                   | 86                     | 70                       |
| Percent of Sample               | 27.6               | 28.0                 | 34.4                   | 28.0                     |

This is supported here by the fact that the number of examinees identified as guessing was higher among those who indicated that English was not their best language. Both tests were more difficult for this subpopulation.

There was a greater number of examinees exhibiting some form of item content/person interaction on the Computation test than on the Logical Relationships test. This probably resulted from the fact that one of the item content areas on the Computation test, application problems, involved considerably more reading and problem solving skills than did the whole number, fraction, or decimal problems. Since reading ability is not generally considered to be part of mathematical ability,

there is the chance that the test may not be unidimensional for those examinees with reading difficulties.

There was very little evidence of plodding on either test, probably due to the fact that ample time limits were set for both tests. Consequently, most students did not run out of time and, therefore did not omit a string of items at the end of the test.

Despite the presence of items that were differentially familiar to males and females on the Logical Relationships test, relatively few examinees were severely affected. There was, however, an interesting content/sex interaction among the examinees who had response patterns that indicated carelessness. Regardless of best language, females tended to exhibit more carelessness on the Logical Relationships test, while males tended to exhibit more carelessness on the Computation test. This seems to fit the stereotypical view of male and female abilities in language and mathematics. The males seemed to be more careful on the Logical Relationships test, perhaps because they do not think of themselves as being as able in that subject. Conversely, the females tended to be more careful on the Computation test, probably for the same reason.

#### Bias Analysis

On the basis of the item calibrations presented earlier, it was possible to identify sets of items that were

differentially familiar to females and to males. Inspection of the words that characterized those items indicated a close conformity to activities that are stereotypically thought of as pertaining to either males or females. Traditional psychometric thinking with respect to bias would suggest that, if a group of items favored a particular subpopulation, all members of the subpopulation would be equally affected by the differential familiarity. The assumption here is that the individual is a random event and can best be described by the population statistics.

The idea unique to Rasch person analysis is that the individual is, in fact, the proper population to study in order to assess the impact of the differentially familiar item on that individual. Due to the mathematical properties of the Rasch model, the analysis of the individual can be freed from the population. It is no longer necessary to think of the person as simply a random sample from the population. Each individual can be studied in his own right independently of the population. Rather than assuming that the degree to which the individual was affected by the presence of biased items was determined by the degree to which the entire population was affected, it is possible to determine the extent to which the individual is affected. Rather than making a correction based on the average effect observed in the population, it is possible to make a correction based solely on the extent to which the individual was affected by the differentially

familiar items.

If one accepts the traditional view of test bias, one may assume that a large number of males and females may be unfairly measured on the items that favor the opposite sex. In order to determine whether the data support this assumption, ability estimates for the three sets of Logical Relationship items were calculated for the 1000 examinees in the person analysis sample. The three sets contained the 11 items that favored males, 10 items that favored females, and the 29 items that favored neither sex. By comparing the ability estimate based on the item set that favored the individual with the ability estimate based on the set of items that favored the opposite sex, it is possible to get an idea of the impact of differentially familiar items on ability estimation. These results are presented in Table 6.

TABLE 6  
FREQUENCY OF HIGHEST ABILITY BY BIASED ITEM SUBSETS  
LOGICAL RELATIONSHIPS TEST

| Sex | English<br>Best<br>Language | Highest Ability Estimate  |      |                             |      |
|-----|-----------------------------|---------------------------|------|-----------------------------|------|
|     |                             | Items That<br>Favor Males |      | Items That<br>Favor Females |      |
|     |                             | N                         | %    | N                           | %    |
| M   | Yes                         | 139                       | 55.6 | 111                         | 44.4 |
| F   | Yes                         | 95                        | 38.0 | 155                         | 62.0 |
| M   | No                          | 148                       | 59.2 | 102                         | 40.8 |
| F   | No                          | 103                       | 41.2 | 147                         | 58.8 |

If the traditional view of test bias is correct, one would expect females to score consistently higher on the items that favored females than on the items that favored males. The same thing should hold for males but in the opposite direction. This expectation is not supported here, since a total of 411 (41.1%) of the males and females analyzed had their highest ability estimates on a set of items that should have been biased against them. Thus, any group correction that might have been considered to correct for the presence of biased items would have been inappropriate for over 40% of the people.

In order to determine if the extent to which the presence of biased items affected individuals on the Computation test, a similar analysis was performed. On this 30-item test 6 items were found to favor males, 7 items were found to favor females, and 17 items were found to favor neither sex. The frequency of higher ability estimates on the items that were biased and those that were not is presented in Table 7.

In this instance 400, (40%) of the examinees had a higher ability estimate on a set of items that favored the opposite sex than on that which that favored their own sex. This is consistent with the results for the Logical Relationships test.

TABLE 7  
FREQUENCY OF HIGHEST ABILITY BY BIASED ITEM SUBSETS  
COMPUTATION TEST

| Sex | English<br>Best<br>Language | Highest Ability Estimate  |      |                             |      |
|-----|-----------------------------|---------------------------|------|-----------------------------|------|
|     |                             | Items That<br>Favor Males |      | Items that<br>Favor Females |      |
|     |                             | N                         | %    | N                           | %    |
| M   | Yes                         | 164                       | 65.6 | 86                          | 34.4 |
| F   | Yes                         | 111                       | 44.4 | 139                         | 55.6 |
| M   | No                          | 160                       | 64.0 | 90                          | 36.0 |
| F   | No                          | 113                       | 45.2 | 137                         | 54.8 |

In this case, however, there appears to be an interaction with the examinee's sex. Females exhibited a higher tendency to score high on items that favored the opposite sex than did males.

#### CONCLUSIONS

From the results presented several conclusions can be drawn.

1) It is possible to identify items that are differentially familiar to various subpopulations by comparing the item difficulties obtained from calibrating the items on the different subpopulations.

2) Given sets of differentially familiar items identified by this technique, it is possible to determine

the degree to which the individual was affected by the presence of these items on the test. The data presented here indicate that, despite the presence of these differentially familiar item sets, relatively few individuals, 3.1% on the Logical Relationships test and 4% on the Computation test, were significantly affected by their presence. These data call into question the traditional view that corrections for biased items should be applied equally for all members of a subpopulation.

3) The data presented here indicate that approximately 1 in every 4 examinees was significantly affected by one or more forms of measurement disturbances. This implies that 1 out of every 4 test scores reported in the standard score metric will not be an accurate reflection of the examinee's true ability. Since the data support several a priori beliefs about measurement disturbances, e.g., that there is more guessing on more difficult tests and that there is more content/person interactions when one of the content areas is measuring a skill normally not considered as part of the variable, it is reasonable to assume that the cases identified are, in fact, instances of measurement disturbance and not outliers of the null distribution.

#### SUMMARY

Depending on the sample from 18% to 34% of the examinees

in this study were significantly affected by some form of measurement disturbance. In a "rights only" scoring procedure, all of these disturbances would have been ignored. The simple sum of the correct responses would have been used as a valid measure of the individual's ability, regardless of the pattern of responses that generated that raw score. It is clear from this analysis that this method is faulty.

With any post hoc analysis it is impossible to determine exactly what caused the measurement disturbance, but the knowledge that the overall test score may not be representative of the examinee's true ability can be extremely useful, particularly if important decisions are being made on the basis of the test results. It should also be noted that person analysis, as described above, is cost-effective. Given existing programs, it is possible to perform person analysis for less than 10 cents per person. This is a small price to pay to insure that the score reported accurately reflects the examinee's true ability.

## REFERENCES

- Draba, R.E. The identification and interpretation of item bias. Research Memorandum No. 26, Statistical Laboratory, Department of Education, University of Chicago, 1977.
- Mead, R.J. Analysis of fit of data to the Rasch model through analysis of residuals. Ph.D. Dissertation, University of Chicago, 1976.
- Mead, R. J. & Kreines, D. C. Person fit with the dichotomous Rasch model. A paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.
- Smith, R.M. An analysis of the individual effects of sex bias. A paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, 1981a.
- Smith, R.M. Person fit analysis with the Rasch model. Qualifying paper, University of Chicago, 1981b.
- Smith, R.M. Detecting measurement disturbances with the Rasch model. Ph.D. Dissertation, University of Chicago, 1982a.
- Smith, R.M. PARM: Person analysis with the Rasch model. Basic computer program, University of Chicago, 1982b.
- Smith, R.M. & Mitchell, V.P. Improving the predictive ability of placement tests using the Rasch model for scoring. A paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, 1979.
- Smith, R.M. & Mitchell, V.P. An analysis of Rasch person fit statistics in college placement tests. A paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. Afterword in Probabilistic models for some intelligence and attainment tests, by G. Rasch. Chicago: University of Chicago Press, 1980.

Wright, B.D., Mead, R.J., & Bell, S.R. BICAL: calibrating items with the Rasch model. Research Memorandum 23B, Statistical Laboratory, Department of Education, University of Chicago, 1979.

Wright, B.D. & Stone, M.H. Best test design. Chicago: MESA Press, 1979.